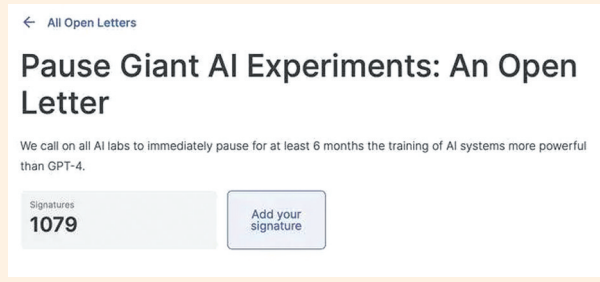
2023年3月30日 星期四
主编 赵本春 责编 李德强
视觉 胡颖 朱兴羽 校审 李勇强

可能导致核战争和更严重疫情，马斯克等千名大佬联名

呼吁暂停训练比GPT-4更强大的AI系统

马斯克

▲马斯克在社交网络上与杰弗里·辛顿和盖瑞·马库斯互动



►未来生命研究所的公开信已有1079人签署

“高级AI可能造成地球生命史上的深刻变化。”这是非盈利机构未来生命研究所（Future of Life）3月29日发布的一封公开信里的警告。这封公开信呼吁人类暂停开发比目前GPT模型更强大的人工智能，为时至少6个月。相关网页显示，已有上千名重量级大佬签名联署，包括特斯拉创始人马斯克、苹果联合创始人沃兹尼亚克等。

其实，最近几天，人工智能风险引发了多位科技领袖的深切担忧。

首先是被称为“人工智能教父”的杰弗里·辛顿，他上周在接受哥伦比亚广播公司采访时表示，人工智能可能发展到对人类构成威胁“并非不可想象”的地步。

随后，AI“大牛”盖瑞·马库斯3月27日发推响应辛顿，28日又发表题为“人工智能风险≠通用人工智能风险”的文章，称超级智能可能会也可能不会迫在眉睫，但在短期内，需要担心“MAI（平庸人工智能）风险”。

27日，推特CEO伊隆·马斯克也加入进来，表达对辛顿和马库斯的赞同。

不过，截至发稿，最受瞩目的OpenAI公司创始人阿尔特曼还没有正式回应这封公开信。

在普通人还在担心AI抢了自己的饭碗时，大佬们似乎看到了更深的一层担忧。

签署这封信的纽约大学教授盖瑞·马库斯说：“这封信措辞并不完美，但精神是正确的：我们需要放慢脚步，直到我们更好地理解AI的后果。AI可能会造成严重伤害，而大公司对他们正在做的事情越来越保密，这使得社会很难抵御可能出现的任何伤害。”

A “加快开发强大的AI治理系统”

未来生命研究所的这封公开信题为“暂停巨型AI实验：一封公开信”，于29日解封。数据显示，公开信已有1079人签署。

信中写道：广泛的研究表明，具有与人类竞争智能的人工智能系统可能对社会和人类构成深远的风险，这一观点得到了顶级人工智能实验室的承认。正如广泛认可的“阿西洛马人工智能原则（Asilomar AI Principles，阿西莫夫的机器人三大法则的扩展版本，于2017年由近千名人工智能和机器人专家签署）中所述，高级AI可能代表地球生命史上的深刻变化，应该以相应的关照和资源进行规划和管理。不幸的是，这种级别的规划和管理并没有发生，尽管最近几个月人工智能实验室陷入了一场失控的竞赛，以开发和部署更强大的数字思维，没有人——甚至他们的创造者——能理解、预测或可靠地控制。

“当代人工智能系统现在在一般任务上变得与人类具有竞争力，我们必须扪心自问：我们是否应该让机器用宣传和谎言充斥我们的信息渠道？我们是否应该自动化所有工作，包括令人满意的工作？我们是否应该发展最终可能比我们更多、更聪明，淘汰并取代我们的非人类思维？我们应该冒险失去对我们文明的控制吗？”这封信写道，“只有当我们确信它们的影响是积极的并且风险是可控的时候，才应该开发强大的人工智能系统。这种信心必须有充分的理由，并随着系统潜在影响的规模而增加。OpenAI最近关于通用人工智能的声明指出，‘在某些时候，在开始训练未来的系统之前进行独立审查可能很重要，并且对于最先进的工作来说，应该同意限制用于创建新模型的计算量增长。’我们同意。那个某些时候就是现在。”

这封信呼吁，所有人工智能实验室立即暂停训练比GPT-4更强大的人工智能系统至少6个月。这种暂停应该是公开的和可验证的，并且包括所有关键参与者。如果不能迅速实施这种暂停，政府应介入并实行暂停。人工智能实验室和独立专家应该利用这次暂停，共同开发和实施一套用于高级人工智能设计和开发的共享安全协议，并由独立的外部专家进行严格审计和监督。这些协议应确保遵守它们的系统是安全的，并且无可置疑。这并不意味着总体上暂停AI开发，只是紧急从奔向不可预测的大型黑盒模型的危险竞赛中收回脚步。

与此同时，信中指出，AI开发人员必须与政策制定者合作，以显著加快开发强大的AI治理系统。至少应包括：建立专门负责AI的有能力的新监管机构；监督和跟踪高性能人工智能系统和大量计算能力；推出标明来源系统和水印系统，以帮助区分真实与合成，并跟踪模型泄漏；强大的审计和认证生态系统；界定人工智能造成的伤害的责任；为人工智能技术安全研究提供强大的公共资金；以资源充足的机构来应对人工智能将造成的巨大经济和政治破坏（尤其是对民主的破坏）。

这封信最后写道，“人类可以享受人工智能带来的繁荣未来。成功创建强大的AI系统后，我们现在可以享受‘AI之夏’，收获回报，设计这些系统以造福所有人，并为社会提供适应的机会。社会已经暂停其他可能对社会造成灾难性影响的技术。我们可以在这个领域也这样做。让我们享受一个漫长的AI之夏，而不是毫无准备地陷入秋天。”

未来生命研究所建立于2014年，是一个非营利组织，由一系列个人和组织资助，使命是引导变革性技术远离极端、大规模的风险，转向造福生活。

截至发稿，这封信已有1079名科技领袖和研究人员签名，除马斯克、辛顿和马库斯外，还包括图灵奖得主约书亚·本希奥、《人工智能：现代方法》作者斯图尔特·罗素、苹果公司联合创始人史蒂夫·沃兹尼亚克、Stability AI首席执行官埃马德·莫斯塔克等科技界领袖人物。

B “可能导致核战争和更严重的疫情”

28日，马库斯在写作平台Substack上写道，一位同事写信给他问道：“这封（公开）信不会造成对即将到来的AGI（通用人工智能）、超级智能等的无端恐惧吗？”

马库斯解释道，“我仍然认为大型语言模型与超级智能或通用人工智能没有太大关系；我仍然认为，与杨立昆（Yann LeCun）一样，大型语言模型是通向通用人工智能道路上的一个‘出口’。我对厄运的设想可能与辛顿或马斯克不同；他们的（据我所知）似乎主要围绕着如果计算机快速而彻底地自我改进会发生什么，我认为这不是一个直接的可能性。”

但是，尽管许多文献将人工智能风险等同于超级智能或通用人工智能风险，但不一定非要超级智能才能造成严重问题。“我现在并不担心‘AGI风险’（我们无法控制的超级智能机器的风险），在短期内我担心的是‘MAI风险’——平庸人工智能（Mediocre AI）是不可靠的（比如必应和GPT-4）但被广泛部署——无论是就使用它的人数而言，还是就世界对软件的访问而言。一家名为Adept.AI的公司刚刚筹集了3.5亿美元来做到这一点，以允许大型语言模型访问几乎所有内容（旨在通过大型语言模型‘增强你在世界上任何软件工具或API上的能力’，尽管它们有明显的幻觉和不可靠倾向）。”

马库斯认为，许多普通人也许智力高于平均水平，但不一定是天才，在整个历史中也制造过各种各样的问题。在许多方面，起关键作用的不是智力而是权力。一个拥有核代码的白痴就可以摧毁世界，只需要适度的情报和不应该得到的访问权限。现在，AI工具已经引起了犯罪分子的兴趣，越来越多地被允许进入人类世界，它们可能造成更大的破坏。

欧洲刑警组织27日发布一份报告，对采用ChatGPT类工具进行犯罪的可能性进行了讨论，令人警醒。“大型语言模型检测和重现语言模式的能力不仅有助于网络钓鱼和在线欺诈，而且通常还可以用来冒充特定个人或群体的讲话风格。这种能力可能会被大规模滥用，以误导潜在受害者将他们的信任交给犯罪分子。”这份报告写道，“除了上述犯罪活动外，ChatGPT的功能还有助于处理恐怖主义、政治宣传和虚假信息领域的许多潜在滥用案件。因此，该模型通常可用于收集更多可能促进恐怖活动的信息，例如恐怖主义融资或匿名文件共享。”

马库斯认为，再加上人工智能生成的大规模宣传，被大型语言模型增强的恐怖主义可能导致核战争，或者导致比新冠病毒更糟糕的病原体的故意传播等。许多人可能会死亡，文明可能会被彻底破坏。也许人类不会真的“从地球上消失”，但事情确实会变得非常糟糕。

“人工智能教父”杰弗里·辛顿在接受哥伦比亚广播公司采访时表示，人工智能可能发展到对人类构成威胁“并非不可想象”的地步。

上周在哥伦比亚广播公司的采访中，“人工智能教父”辛顿被问及人工智能“消灭人类”的可能性时说，“我认为这并非不可想象。我只想说这些。”

马库斯指出，相比辛顿和马斯克担忧天网和机器人接管世界，应该更多思考眼前的风险，包括恐怖分子在内的犯罪分子可能对大型语言模型做什么，以及如何阻止他们。

据澎湃新闻、红星新闻等